Program on Education Policy and Governance Working Papers Series

Teacher Incentive Pay and Educational Outcomes: Evidence from the NYC Bonus Program

Sarena Goodman
Columbia University

Lesley Turner
Columbia University

PEPG 10-07

Prepared for the PEPG Conference
Merit Pay: Will It Work? Is It Politically Viable?

Harvard Kennedy School
Cambridge, Massachusetts
June 3-4, 2010

Teacher Incentive Pay and Educational Outcomes: Evidence from the NYC Bonus Program

Sarena Goodman
Columbia University

Lesley Turner
Columbia University

May 2010*

Abstract

Teacher compensation schemes are often criticized for lacking a performance-based component. Proponents of merit pay argue that linking teacher salaries to student achievement will incentivize teachers to focus on raising student achievement and stimulate innovation across the school system as a whole. In this paper, we utilize a policy experiment conducted in the New York City public school system to explore the effects of one performance-based bonus scheme. We investigate potential impacts of group-based incentive pay over two academic years (2007-2008 and 2008-2009) on a range of outcomes including: teacher effort, student performance in math and reading, and classroom activities, measured through environmental surveys of teachers and students. We also explore impacts on the market for teachers by examining teacher turnover and the qualifications of newly hired teachers. Overall, we find the bonus program had little impact on any of these outcomes. We argue that the lack of bonus program impacts can be explained by the structure of the bonus program. Group bonuses led to free-riding, which significantly reduced the program's incentives. Once we account for free-riding, we find evidence that the program led teachers to increase their effort through a significant reduction in absenteeism. When considering the effectiveness of performance-based teacher pay, the structure of incentives matter.

* Correspondence should be sent to ljt2110@columbia.edu. We are especially grateful to Jonah Rockoff for his thoughtful comments and advice. We would like to also thank Todd Kumler, Bentley MacLeod, Ben Marx, Petra Persson, Jesse Rothstein, Miguel Urquiola, Till Von Wachter, and Reed Walker and seminar participants at the Columbia applied microeconomics colloquium, AEFA annual meeting, and Teacher's College Economics of Education Workshop for helpful feedback. We are grateful to the New York City Department of Education for the data employed in this paper.

1. Introduction

Teacher compensation schemes are often criticized for their lack of performance pay and relatively low pay in general. These features of teachers' salaries potentially lead to sorting and adverse selection in the market for teachers, as well as inefficiently low effort provided by existing teachers. A large body of empirical research shows that in other sectors, incentive pay successfully increases worker effort and output.[1] Thus, proponents of teacher merit pay argue that linking teacher salaries to student achievement will induce teachers to focus on raising student achievement and stimulate innovation in the school system as a whole.[2] Performance pay is most effective when employers can measure and reward on-the-job performance. However, education is a complex good; it is difficult to observe and appropriately monitor the behavior of educators and their respective contributions to the production of education since production depends not only on a student's current teacher but on the effort provided by past teachers. Thus, in theory, while performance pay should improve educational outcomes, the structure of teacher incentive pay will matter.

In this paper, we investigate the impact of group-based teacher incentive pay on teacher effort, student achievement, and teacher sorting across schools. We take advantage of a policy experiment conducted in New York City. In the fall of 2007, 181 schools were randomly selected from a group of high-poverty schools. These schools were eligible to earn school-level bonuses based primarily on student achievement on state math and reading exams.[3] Potential bonus payments represented between three to seven percent increases in teachers' annual pay.

We examine the impact of incentive pay on a wide range of outcomes including teacher effort, measured by absenteeism, student achievement in math and reading, measured by performance on New York state exams, and outcomes from surveys of teachers and students, including classroom activities and school-level policies. Since bonus payments are based the performance of the school as a whole, teachers only earn bonuses if school-wide targets are met. Thus, a teacher's ability to affect the probability of receiving a payment is decreasing in the number of teachers with tested students, indicating the potential for free-riding as the number of

[1] Gibbons (1998), Lazear and Oyer (2010), and Oyer and Schaefer (2010) review this literature.
[2] These compensation schemes are generally most effective in sales jobs and those that involve operating machines. Macleod and Parent (1999) provide an overview of other sectors that employ incentive-based pay schemes.
[3] The program also included 39 secondary schools. Since bonus receipt for high schools was based on different outcomes for high schools, we focus on elementary and middle schools and schools serving children in kindergarten through 8th grade (K-8 schools).

teachers grows large. To test whether free-riding dilutes bonus program incentives, we examine whether the program impacts vary with the number of teachers with students who are tested (and therefore contribute to the probability that a school qualifies for the bonus award). Finally, to determine whether the program increased relatively disadvantaged schools' ability to recruit or retain qualified teachers, we test whether eligibility to earn bonuses affected 1) end-of-year teacher turnover or 2) the quality composition of entering classes of teachers.

We find some evidence that teachers responded to the program by increasing effort once we account for heterogeneity in the potential for free-riding. Specifically, in schools with a small number of teachers, we find the program led to significant increases in teacher effort. However, increases in attendance are not large enough to translate into test score gains and, not surprisingly, we find little effect of the bonus program on student achievement in the first or second year of the program. We find no discernable effect on in-class or school-wide policies reported by students and teachers, such as additional tutoring sessions or increased use of student achievement data. Finally, we show that the bonus program had little effect on teacher turnover or the qualifications of newly hired teachers. The first section of our paper describes the bonus program. Section 3 discusses the difficulties and theoretical implications associated with implementing merit-based pay in schools. Section 4 provides an overview of the data, section 5 outlines our empirical framework and presents results, and section 6 concludes.

2. The New York City School-Wide Bonus Program

We use a policy experiment implemented by the New York City Department of Education (DOE) in the fall of 2007, the "School-Wide Bonus Program" (hereafter, the bonus program). Both the DOE and the United Federation of Teachers (UFT) endorsed the program and it was lauded as an innovative model for teacher performance pay. In November 2007, 181 schools serving kindergarten through eighth grade were randomly selected from a group of schools designated as "high need"; 128 schools were assigned to the treatment group. Treatment schools were eligible to participate in the program, contingent on teacher approval: 55 percent of full-time a school's United Federal of Teachers (UFT) staff had to vote in favor of the program. Twenty-five schools voted to not participate or withdrew from the treatment group prior to a vote. Treatment schools that voted in favor of the program would earn a lump-sum bonus if school-wide goals, based primarily on student achievement, were met. Schools that either

achieved a target score or were awarded an "A" accountability grade (explained below) for two consecutive years received bonuses equal to $3,000 per union teacher, while schools that fell short but managed to meet 75 percent of the target score received $1,500 per union teacher. Schools that did not achieve their target did not face consequences beyond the absence of bonus pay.

Each participating school formed a four-member compensation committee, consisting of the principal, a second administrator, and two union representatives elected by the school's UFT members.[4] This committee was required to submit a bonus distribution scheme *after* student math and reading exams. Thus, the ultimate split of the bonus award should not affect an individual teacher's effort decision in the first year of the program. Bonus program guidelines stipulated that all union teachers receive a bonus payment; the committee chose bonus amounts and whether non-union employees also received funds. The committee was unconstrained in choosing a distribution plan except that bonuses could not be explicitly based on tenure. Around half of treatment schools choose an approximately equal distribution (e.g., the difference between the highest and lowest bonus payment was less than $100), while in the remainder of schools, the difference between the highest and lowest bonus ranged from $200 to $5000 (Figure 1).[5] In schools that choose an equal distribution of bonus payment, the full $3,000 award represents a seven percent increase in the salary of teachers at the bottom of the pay scale and a three percent increase for the most experienced teachers.[6]

The timing of program announcement and the selection of schools into the treatment group did not allow much room for behavioral responses to the program in its first year. The school vote took place in November 2007, only one month before reading exams were taken in January and three months before the math exams in March. However, the program was continued in the 2008-2009 school year and all schools in the original treatment group voted to participate in the second year of the program. Of the 158 treatment schools that voted to participate in the first year of the program, 89 (56 percent) received bonus payments. The bonus pool averaged $160,095 per school, and amounted to a total of $14.2 million district wide in the first year. In the second year of the program, the vast majority (91 percent) of treatment schools earned bonus awards, totaling $27.1 million.

[4] See http://www.uft.org/member/rights/bonus/moa/.
[5] In schools that choose unequal distributions, on average, a standard deviation of teacher bonuses was $143.
[6] Teacher salary schedules are available at http://www.uft.org/member/contracts/moa/salary_schedules

The 2007-2008 school year also marked the implementation of the DOE's new accountability system. Under this system, schools received progress reports with accountability grades designed to summarize a school's overall performance on a multidimensional metric of student learning.[7] Each school's performance was scored relative to the entire district and to a group of "peer schools," which included the 40 closest schools according to a "peer index" measuring student demographic characteristics and prior year test scores.[8] Each school's progress report documented its score on this metric, the corresponding accountability grade, and a target score. Schools that received lower accountability grades needed to make large improvements to reach their target scores. Although the accountability system was more complex than systems based on a single metric (e.g., the percentage of students achieving proficiency), teachers and administrators received training on how to interpret the complicated set of measures determining a school's grade. Rockoff and Turner (forthcoming) find that receiving an F or D led to a significant improvement in student test scores, providing some evidence that school employees understood that performance under the accountability system was dependent on student achievement.

The details of the accountability system are important for our analysis: schools were selected into the experimental sample based on their peer indices, and treatment schools had to reach their target scores to qualify for teacher bonuses. Furthermore, the accountability system provided additional incentives to schools participating in the bonus program. Schools that earned an A or B accountability grade received rewards (e.g., principal bonuses, additional funds based on students transferring from schools receiving a poor grade), while schools that received D and F grades faced consequences (e.g., risk of school closure and removal of principal). It is important to note that our results estimate the effect of group-based teacher performance pay in a district where schools are already under accountability pressure.

[7] The metric was calculated from a measure of school environment (student attendance and results from survey of parents, teachers, and students), student performance (average student achievement on reading and math exams, median proficiency, and percentage students achieving proficiency), student progress (average change and percent making progress on math and reading exams), with the option of an extra credit for exemplary student progress among high-need students.

[8] For elementary schools and those serving kindergarten through eighth grade (K-8), the index was based on a function of the percentage of students that were English language learner (ELL), special education, Title I free lunch, and minority. For middle schools, the peer index was based on the 4th grade reading and math test scores of current students. These different constructions actually encapsulate consistent metrics for relative disadvantage, as the components for the elementary/K-8 peer index are very strong predictors of 4th grade test scores. Therefore, the two methods should yield reasonably close measures.

3. Incentive Pay and Teacher Effort

Properly-structured performance pay can offset shirking behavior and encourage employees to provide costly effort.[9] Allowing compensation to vary with performance also aligns worker and employer incentives, providing information about the most valued aspects of an employee's job. When a job involves several tasks or when the nature of such tasks are broadly defined, incentive pay can help resolve confusion as to how best to fulfill responsibilities. If in, at least some public schools, teachers exert an inefficiently-low amount of effort, or focus their effort on tasks with low marginal returns, teacher incentive pay may lead to increases in student learning. In the long-run, a performance-based element of teacher pay may combat wage compression in the profession and increase the ability of individuals opting into the teaching profession (Hoxby and Leigh, 2005).

There are several reasons why performance pay in the educational sector may not be as effective as it is for sales or other output-based jobs. Performance-based compensation is only feasible when reasonable measures of inputs or output are available. Output must be at least partially contingent on worker effort for incentive pay to increase productivity. Thus, performance pay is most effective when employers can quantify worker effort or when measurable output is clearly linked to effort provided. For example, in piece-rate jobs, the quality and amount of product attributable to one employee is readily observable. However, it is costly to monitor teachers and difficult to measure teacher effort. Educational goals are difficult to define and the desired output is hard to measure; longer-term outcomes, such as future wages, are arguably the most important, but it infeasible to tie current teacher salaries to outcomes that are not observed in the short-run. As a result, teacher incentive pay is generally linked to outcomes that are correlated with future wages and also easy to measure, such as student performance on standardized tests. Second, academic performance is multifaceted and difficult to attribute to a particular source; for instance, test scores depend on a student's current teacher and also upon the effort of prior teachers. Finally, public education is not provided in a fully competitive market; thus, there are no market-based standards for teacher performance. Administrators are often left to define their own parameters of teacher performance, which are

[9] Effort extraction is just one motivation for incentive-based pay. Incentive systems are also used to improve sorting of workers across jobs and to select quantity versus quality of output (Lazear, 1986).

likely to be tailored toward own school demands but may be inconsistent with broader societal objectives.[10]

Education is a complex good. Teachers must complete multidimensional tasks and must allocate their effort across several activities. Holmstrom and Milgrom (1991) demonstrate that the performance metric to which compensation is tied affects how effort is distributed across duties. When production involves a multi-dimensional task and incentives are provided focus on a single dimension, workers optimally expend less effort on other tasks. Thus, the performance measure used to evaluate performance and teachers' potential responses are both important to consider. Although test scores are easily measured, tying performance pay to testing outcomes may incentivize teachers to focus on narrowly-defined basic skills that appear on exams (e.g., "teaching to the test") or overtly manipulate test scores (e.g., Levitt and Jacob, 2003; Jacob, 2005; Figlio, 2006; Figlio and Getzler, 2006; Cullen and Reback, 2006).[11] On the other hand, teachers might respond to incentive payments by increasing effort along several margins; for instance, showing up to school more or increasing time with students outside of the classroom through extra-help sessions and after-school tutoring. It need not be the case that these teaching activities immediately translate into higher test scores. Rather, it is only necessary that teachers believe these behaviors are correlated with student achievement.

In practice, the structure of teacher performance pay varies significantly. Current systems include tournaments, where the teachers whose classrooms experience the greatest gains receive award, to bonuses tied to fixed achievement thresholds, and from individual incentives to group-based incentives, where bonus payments are contingent on school- or district-wide performance. The specifics of how awards are allocated, the size of potential bonuses, and the metrics on which bonuses are based are all important.

Figlio and Kenny (2007) document a positive cross-sectional relationship between individual-based teacher performance pay and student achievement in the United States. The most effective systems appear to be those where awards were difficult to earn and only a small number of teachers received incentive payments. However, these results are confounded by the

[10] In contrast, Ballou and Podgursky (1997) have found that private schools are more likely to rely on teacher performance incentives.

[11] For instance, teachers might change exam responses, give answers to students, or distribute exam questions before the test date. Teachers may exempt particular students from the test, either by encouraging or forcing them to miss school on testing days, or by reassigning them to special status (e.g., special education classes) that allow them to either bypass the exam completely, receive more time for an exam, or give schools the ability to reweight their scores.

possibility that better schools might be more willing to adopt bonus pay, so the authors remain agnostic regarding the direction of causation. Atkinson et al. (2009) examine performance pay in England using quasi-experimental variation the implementation of a performance based-scheme and find positive impacts on student test scores and teacher value-added. Experimental evidence from a paper on individual teacher incentives in Israel is consistent with these findings (Lavy, 2009). Teachers were eligible for cash prizes for their students' relative performance. Incentive payments, ranging from 6 to 30 percent of teachers' average annual salary, led to an increase in both the proportion of students taking a high school exit exam and the performance among test-takers. These student achievement gains likely stemmed from an increase in after-school sessions, evidence of increased teacher effort in response to potential rewards.

There is less evidence on the effectiveness of group-based teacher incentives. In theory, group incentive payments will be the most effective when the production technology is truly joint. If an individual teacher's effort has a positive effect on the effort chosen by other teachers, then group incentives are optimal (Itoh, 1991). Jackson and Bruegmann (2009) find positive spillovers associated with the presence of effective teachers. In the case of group incentives, when the return to increasing effort for any one member is diluted to the point where the marginal cost of effort exceeds the expected bonus, free-riding among team members may result. Ahn (2009) finds evidence of free-riding among teachers in a system involving group bonuses.

Lavy (2002) shows that incentive payments based on school-wide performance increased student test scores and participation on matriculation exams in Israel, but the percentage of students who received matriculation certificates, arguably the longer-run outcome of interest, was not affected. Glewwe et al. (2003) examine the effects of a school-based teacher incentive experiment in rural Kenya, where teachers in grades 4 to 8 were eligible for prizes based on their school's relative performance on annual district exams. In winning schools, all teachers in these grades received equal bonuses. The authors find short-term improvements in test scores but no long-term gains, potentially evidence of gaming. Finally, with a randomized experiment in India where schools were selected to be eligible for either individual or group piece-rate payments based on improvements in student test scores, Muralidharan and Sundararaman (2009) find a large positive impact, across all grades, districts, and student competency levels, for both group and individual teacher incentives, although the longer-run impacts of individual teacher incentives are largest. Most encouragingly, they find that incentive pay led to improvements in

both mechanical and conceptual areas of achievement and in subjects that were not linked to incentive pay.

4. Data and Descriptive Results

We obtained a list of treatment and control schools from the DOE. Our analyses focus on schools classified as elementary, middle, and K-8 (schools serving kindergarten through 8th grade). Schools were selected into the experimental sample (i.e., treatment and control schools) based a "peer index" derived from student characteristics and academic achievement. Schools falling below a certain peer index cut-off were eligible for random selection into the treatment and control groups.[12] A total of 181 schools were originally selected into the treatment group. This group includes 25 schools that either voted to abstain (23 schools) or withdrew from prior to the vote taking place. The control group included 128 schools.

The majority of our outcome variables are drawn from publicly-available DOE data.[13] To create measures of academic achievement, we calculate average math and reading test scores for each school for the 2006-2007, 2007-2008, and 2008-2009 school years (hereafter 2007, 2008, and 2009 school years). From this data, we also construct a measure of the percentage of students classified as proficient in each subject. We take advantage of school-level results from annual surveys of teachers and students conducted by the DOE as part of the accountability system.[14] Specifically, we use the questions from the student survey on the extent to which: 1) students completed essays and research projects and 2) classroom activities including group work, class discussions, and "hands-on activities such as science experiments." We also measure the availability of tutoring, using questions on whether tutoring was offered before or after school. From the teacher survey, we use a question asking whether teachers use student achievement data, such as students' test results from prior years or "periodic examinations" during the school year, to inform their lesson planning. We also create a measure of whether teachers believed students faced high standards and expectations. Finally, in some specifications,

[12] A small number of schools initially belonging to the experimental sample were excluded prior to random assignment. The exclusion of these schools will not affect the internal validity of our results. We examine the characteristics of these schools to determine if the external validity of our results is compromised, and find little differences between these schools and the final experimental sample (results available upon request).
[13] See http://schools.nyc.gov/Accountability/data/default.htm for details (accessed 4/25/2010).
[14] Available at http://schools.nyc.gov/Accountability/tools/survey/default.htm (accessed 4/25/2010).

we include information on each schools performance under the new NYC accountability system, including each school's accountability score and peer index.[15]

We aggregate data from individual teachers to test whether the bonus program had an effect on teacher absenteeism and whether the number of teachers within a school dilutes the incentives of the bonus program, in line with theoretical predictions for individual behavior in the presence of free-riding opportunities. We also use this data to create measures of teacher turnover or the characteristics of newly-hired teachers. We aggregate student-level data to create measures of the percentage of students in each school that are English Language Learners (ELL), special education students, Title I free lunch recipients, and minorities.

We first compare the characteristics of the 209 schools in the experimental sample to other schools in New York City. We restrict our universe to the 987 schools serving students in kindergarten through eighth grade that received accountability grades and were not charter schools or schools that only serve special education students. Given that schools with peer indices at the bottom of the distribution were selected into the bonus program, it is not surprising that the experimental sample differed from the remainder of NYC schools across a number of dimensions (Table 1). On average, schools in the experimental sample had a higher proportion of English Language Learners (ELL), special education, minority students, and students eligible for the Title I free lunch program, as well as lower average math and reading scores. Teachers in the experimental sample had slightly less experience and almost twice as many absences than teachers in other New York City schools. Finally, experimental schools had lower enrollment and fewer teachers than other schools.

4.1 Was Randomization Successful?

Our ability to take advantage of random assignment in making causal inference about the effects of teacher incentive pay depends on the success of random assignment. If random assignment was successful, the observable characteristics of treatment and control group schools should be similar. Table 1 compares the characteristics of treatment and control schools prior to selection into the treatment group, where the group of treatment schools includes all schools selected to participate in the bonus program, including the 25 schools which voted to not take

[15] Middle schools and elementary/K-8 schools have different metrics underlying their respective peer indices that also have different scales. Thus, for descriptive purposes, we standardize each type of school's peer index to have a mean of zero and standard deviation of 1.

part in the program.[16] Treatment and control schools are similar in terms of enrollment, accountability outcomes, student demographics, and teacher characteristics. We find no significant differences between the observable characteristics of treatment and control schools.

5. Regression Framework and Results

To see the advantage of a randomized experiment in estimating the effect of teacher incentives, consider the following model:

$$(1) \qquad Y_{jt} = \delta D_{jt} + \boldsymbol{X_{jt}\beta} + \varepsilon_{jt},$$

where Y_{jt} is the outcome of interest for school j in year t (for example, average math scores in 2008), $\boldsymbol{X_{jt}}$ is a vector of school characteristics in year t, ε_{jt} is a stochastic error component, and D_{jt} is an indicator variable for whether the teachers within the school are eligible for bonus payments. For δ to have a causal interpretation in the absence of random assignment, $\boldsymbol{X_{jt}}$ must include all factors that are correlated with Y_{jt}, including whether a school received the treatment, observable components like average student demographic and other characteristics, and unobservable components such as family history and students' cognitive abilities. In general, data limitations will prevent us from adequately accounting for all relevant school and student characteristics. If these omitted variables are correlated with the included variables, then the estimated parameter will be biased.

However, since selection into the teacher incentive program is determined by random assignment, it is independent of the omitted variables. Thus, with random assignment, a comparison of mean outcomes in treatment and control schools should yield an unbiased estimate of the effect of treatment on the outcomes of interest. The identifying assumption requires that there be no contemporaneous shock that affects the relative outcomes of the treatment schools in the same period as the treatment. Such a shock would be highly unlikely in our setting given the randomized nature of the treatment.

Our primary regression specification takes the following form:

$$(2) \qquad Y_{jt} = \delta D_{jt} + \varepsilon_{jt}$$

[16] Appendix Table A1 compares the characteristics of treatment schools by whether or not they voted to participate in the program. Schools voting "no" are largely similar to schools that received the treatment, although, on average, these 25 schools were relatively less disadvantaged and their students had higher test scores.

where $D_{jt} = 1$ if a school was *eligible* for the bonus program (regardless of whether the school ultimately choose to participate). In the program evaluation literature, these results are referred to as intent-to-treat estimates. We estimate the equation with ordinary least squares, where school observations are weighted by the group size (e.g., number of students tested when the dependent variable is average math scores, number of teacher survey respondents for teacher survey outcomes). Although with successful randomization, this approach should estimate the true effect of the bonus program, we estimate a second specification that includes a vector of control variables to reduce residual variance. These controls include the outcome in the year prior to the intervention, to address any baseline differences between treatment and control schools, indicators for school type (i.e., elementary, middle, or K-8), demographic composition (i.e., percentage of students that are ELL, special education, free lunch, and minority), peer index, and accountability score (since this score determines a school's target score). In our final specification, we instrument for participation in the bonus program with a school's original assignment using two-stage least squares. These estimates can be interpreted as the impact of the treatment-on-the-treated.

5.1 Teacher Effort and the Free-Rider Problem

We first examine whether teachers increased their effort in response to the bonus program. In theory, teachers should respond if the expected marginal benefit is greater than the marginal cost. Although we do not directly observe effort, we can measure teacher attendance. Absences are more common among teachers than in other sectors and absenteeism has been shown to have a negative effect on student achievement (Clotfelter et al., 2009; Miller et al., 2008). Using data on absences among New York City teachers, Herrmann and Rockoff (2009) estimate that an additional 10 absences leads to a 0.01 standard deviation reduction in test scores.

We run a series of regressions where the dependent variable is average absences between the months of November 2007, when schools first learned of their eligibility for the bonus program, and March 2008, when the last exams were taken. If teachers were uncertain of whether the bonus program would continue for more than one year, changes in behavior should be largest over this period. Table 2 presents these results. Each cell contains the estimates from separate regressions of the effect of the bonus program on the number of absences per teacher. The second specification controls for prior year absences and other school characteristics, while the third specification instruments for program participation with initial random assignment. We

separately examine absences among teachers with tested students (e.g., teachers for grades 3 through 5 in elementary schools and math and reading teachers in middle schools). We only consider absences that teachers have some control over – those taken for illness and personal business, excluding days missed due to death in the family, injury, jury duty, absences required by the school system (e.g., for professional development activities), conference attendance, and religious holidays. The first three columns of Table 2 show the bonus program had no effect absences, both among all teachers and for teachers with tested students.

However, the probability that a treatment school reaches its goal and receives a bonus award depends largely on student performance on math and reading exams. Thus, any incentive for teachers to increase effort is decreasing as the number of teachers with tested students grows large.[17] Consider two extremes, a school with only one teacher with tested students and a school with an infinite number of such teachers. In the first case, the teacher will either choose to increase her effort to the level necessary to achieve the school's goal or not respond (if the size of the bonus is less than the cost of exerting this level of effort). In the second case, each individual teacher has no ability to determine whether the school receives a payment and will optimally not respond to the bonus program. Thus, we examine whether treatment effects are related to the number of teachers with tested students.[18] We first de-mean our measure of the number of such teachers, and then include an interaction with the treatment indicator; the point estimate for the treatment indicator denotes treatment effects for the school with an average number of teachers with tested students (columns (4) through (6)). A negative coefficient on the interaction between number of teachers and treatment would provide evidence that the program impacts are diluted by free-riding.

The interaction term is negative and marginally significant, providing suggestive evidence that the bonus program incentives are diluted in schools where the potential for free-riding is large. These results suggest that for schools with fewer than 5 teachers in tested classrooms (schools in the lowest decile in terms of number of teachers), the bonus program increased attendance by 0.5 days per teacher, which translates into 2.5 fewer absences over the five month period we examine. Considering the estimates of Herrmann and Rockoff (2009), this

[17] Additionally, monitoring may be easier in schools with fewer teachers, counteracting free-riding incentives (Holmstrom, 1982).

[18] A small number of middle and K-8 schools do not have information on the number of teachers teaching tested subjects, thus, these schools are not included in regressions where the dependent variable is absences among teachers with tested students.

reduction in absenteeism would only lead to a small (0.003 standard deviation) improvement in test scores.

School compensation committees had the potential to mitigate the free-riding problem, especially if teachers believed that their portion of the bonus would be based on their contribution to the school's goals. Unfortunately, we do not have information on which teachers were designated to receive larger or smaller bonuses in schools that differentiated payments. Schools that choose an unequal distribution had a slightly larger portion of teachers with tested students than those that choose to equal distribute potential bonuses, but these differences are not significant. When we examine the characteristics of treatment schools according to whether bonuses were equally distributed, we find no differences in characteristics except for teachers' survey responses. Teachers in schools that choose an unequal distribution consistently gave their school lower ratings, on survey questions ranging from academic achievement to communication and cooperation (results available upon request). One interpretation of this finding is that in less cohesive schools, the compensation committee believed that free-riding would be a larger concern.

5.2 Student Math and Reading Achievement

To preview our estimates of the impact of the bonus program on student achievement, Figures 2 and 3 display the distribution of average math and reading scores within treatment and control schools in 2007, 2008, and 2009. On average, all NYC schools experienced an increase in average student performance in the two years following the implementation of the program; this pattern holds in the experimental sample. If the bonus program had an impact on test scores, we should observe a shift in the distribution among treatment schools, relative to control schools. However, there are no significant differences in the distribution of test scores in either subject in 2008 or 2009.

Table 3, which displays results from regression estimating the impact of the program on average math and reading exam scores, confirms these findings. We do find any significance impact on aggregate school performance in the first year of the program. The point estimates are negative and quite small, given that the student level standard deviations in 2008 of reading and math scores are 35 points and 40 points, respectively.[19] Even if teachers did respond to the

[19] Winters (2009) estimates the impacts of the bonus program on test scores at the student level and finds similar results.

program, one might not expect to observe any effects in the initial year, given how closely the tests followed program implementation. However, eligibility to earn bonuses did not have any effect on student achievement in 2009 (Table 3, columns (4) through (6)), and, if anything, these results suggest the program may have had a negative impact on student achievement.[20] While, we do find that the bonus program led to a significant reduction in the percentage of students classified as proficient in math in 2009, the magnitude of this effect – approximately a 1.5 percent reduction in proficiency – is quite small (Table 4).[21]

In Table 5, we test for evidence of free-riding and allow treatment effects on math and reading scores to vary by the number of math and reading teachers, respectively.[22] We also find evidence of free-riding, although the estimated effects are small in magnitude. In schools at the bottom of the distribution of teachers, we estimate positive but insignificant effects of the bonus program. Even in schools where the bonus program approximated individual incentive pay system, the size of potential awards was not sufficient to induce large enough increases in effort to affect student achievement.

5.3 Heterogeneity in Bonus Program Impacts

Although the bonus program had little effect on average student achievement, tying bonuses to the structure of the NYC accountability system provided incentives for schools to focus on students at along different points of the achievement distribution. In line with recent research examining the effect of accountability systems on performance among different student subgroups (Cullen and Reback, 2002; Figlio and Getzler, 2002; Figlio, 2006), we test whether the bonus program had heterogeneous impacts across student subgroups.[23] We find little difference in the impact of the bonus program across different types of students (results available

[20] Four schools in the treatment group were closed at the end of the 2008 school year, thus, our sample decreases by four in the second set of regressions. Additionally, three schools did not receive an accountability score in 2008, thus, these schools are dropped in columns (5) and (6). Our 2008 results remain unchanged when we restrict the sample to exclude these schools.

[21] Students are considered proficient if they achieve a set score on the state exams and are considered to be meeting learning standards.

[22] In elementary schools and lower grades of K-8 schools, these teachers also teach other subjects.

[23] Unlike other accountability systems (e.g., No Child Left Behind) that depend on the proportion of students that reach an absolute level of proficiency, the NYC accountability system contains incentives to focus on some groups of students. While the accountability system awards "points" for average school performance and changes in performance for individual students, students with certain characteristics may be double or even triple counted: those whose prior-year achievement placed them in the lowest third of their grade, students on the cusp of proficiency and those close to the school median, and ELL and special education students also are given more weight.

upon request). An additional concern is that schools might also respond to the bonus program by removing students from the test-taking pool or reclassifying higher performing students as either ELL or special education to take advantage of the increased weight placed on these students' achievement. We do not find that the proportion tested students classified as ELL or special education within treatment schools increased relative to control schools (Appendix Table A2).

Treatment schools face different incentives according to their accountability grades. Since schools that received an A on their progress report needed only to maintain this grade, the bonus program may not have provided a large incentive to teachers in treatment schools to alter their behavior. Conversely, both treatment and control schools receiving low grades had additional motivation to improve student test scores, as they faced school closure or principal removal if student achievement did not improve in the following year. Schools in the middle of the grade distribution perhaps faced the strongest difference in incentives. Thus, we also test whether treatment effects vary along this dimension, grouping schools into three separate bins by their accountability grades: A, B/C, and D/F. Our estimates become noisy, likely due to the small sample size within each grade-grouping, but are consistent with our main results. We find no significant differences in treatment effects between these grade groupings or for schools at the center of the grade distribution where the difference in incentives between treatment and control schools is greatest (Appendix Table A3).

Finally, we test the receipt of bonus payments had an effect on student achievement in the second year of the program. Since we know the metric used to determine which schools received bonuses, we simulate bonus receipt in the control group and interact eligibility for the bonus program (or treatment in our 2SLS specifications) with predicted bonus receipt. As shown in Appendix Table A4, we do not observe any heterogeneity in 2009 student achievement by bonus receipt.

5.4 Student and Teacher Survey Results

It is possible that teachers and school administrators responded to the bonus program, but that these behavioral changes did not translate into increased student achievement. We might also be concerned that incentives to specifically focus on student achievement would lead to a reduction in other classroom activities (Holmstrom and Milgrom, 1991). Thus, we explore whether the bonus program led to changes in teacher behavior and school policies using results

from the DOE's annual surveys of teachers and students.[24] We test whether the program induced any changes in classroom activities, by examining the extent to which students reported working on "essays or projects" and "group work or hands-on activities". We also test whether the program increased opportunities for before- or after-school tutoring sessions. Only students in grades six or higher completed the environmental survey, thus, we lose a number of schools, mostly at the elementary level. We do not find significant effects of treatment on student reports of participating in group or hands-on learning activities or on whether they completed projects or essays in class, although both of these outcomes are positively correlated with treatment and in the third specification, the latter measure comes close to conventional significance levels (Table 6, Panel A). Additionally, the bonus program appears to have no significant impact on the availability of tutoring.

Although the bonus program targets teachers, one might also expect it to induce changes in school-wide decisions. However, we do not find evidence of institutional responses to the intervention (Table 6, Panel B). There are no significant treatment effects on teachers' use of student data and the point estimates of the impact of treatment on these outcomes are small. The second measure we examine from the teacher– whether teachers believed students in their school were held to high expectations – is negative and approaches conventional significance levels in the second and third specifications. These results provide little evidence that teachers substituted award from more complicated activities in favor of test prep.

5.5 Teacher Characteristics and Turnover

Finally, we investigate whether the program bonus program led to changes in the quality of new teachers and reduced teacher turnover, in line with literature on sorting (e.g., Clotfelter et al., 2006). Poor schools traditionally have more difficulty hiring and retaining highly-qualified teachers (Hanushek and Rivkin, 2007). If the bonus program increased the supply of qualified teachers willing to work at treatment schools, any resulting student achievement gains will lag these changes by at least a year.

We first examine whether the bonus program led to a reduction in teacher turnover. In a given year, approximately 10 percent of NYC teachers leave the city while an additional 8 percent switch schools within the city. As shown in Panel A of Table 7, the bonus program did

[24] For ease of interpreting results, all survey outcomes are standardized to have a mean of zero and standard deviation of 1 across all NYC schools, according to school type.

not reduce either type of turnover. Second, we test whether treatment schools experienced an increase in the qualifications of newly hired teachers (Table 7, Panel B). Recent studies have found Teach for America volunteers to be more effective in raising test scores than regularly-certified new teachers (Decker et al., 2004; Kane et al., 2008). However, we find little effect on the proportion of new teachers hired through Teach for America. We also test whether the bonus program increased the proportion of new hires holding a masters degree and find positive but insignificant treatment effects.

7 Conclusion

In general, empirical research shows that incentive pay enhances effort, output, and other desirable outcomes. However, despite significant expenditures on the NYC bonus program, we find little effect on student achievement. We estimate a small reduction in absences for teachers with the largest incentives, suggesting a limited effect on teacher effort. However, increases in effort were not large enough to translate into increased student achievement, even in schools where bonus payments approximated individual incentives. We find no significant impact on student or teacher assessments of classroom activities, tutoring, or administrative decisions. Finally, it does not appear that the program affected teacher turnover or the quality of new teachers within treatment schools. The fact that we find a response among teachers with the largest incentives provides evidence that teachers did understand the bonus scheme, but the potential for free-riding in most schools reduced the incentives for teachers to increase their effort. Our results underscore the fact that the structure of teacher performance pay is important.

References

Ahn, T. (2009) "The Missing Link: Estimating the Impact of Incentives on Effort and Effort on Production Using Teacher Accountability Legislation," University of Kentucky.

Atkinson, A., Burgess, S., Croxson, B., Gregg, P., Propper, C., Slater, H., and Wilson, D. (2009) "Evaluating the Impact of Performance-Related Pay for Teachers in England," Labour Economics, 16: 251-261.

Ballou, D. (2001) "Pay for Performance in Public and Private Schools," Economics of Education Review, 20(1): 51-61.

Clotfelter, C., Ladd, H., and Vigdor, J. (2009) "Are Teacher Absences Worth Worrying About in the U.S.?" Education Finance and Policy, 4(2): 115-149.

Clotfelter, C., Ladd, H. and Vigdor, J. (2006) "Teacher-Student Matching and the Assessment of Teacher Effectiveness," National Bureau of Economic Research working paper #11936.

Cullen, J. B. and Reback, R. (2006) "Tinkering Toward Accolades: School Gaming Under a Performance Accountability System," In T. Gronberg and D. Jansen (Eds), Advances in Applied Microeconomics, 14.

Decker, P., Mayer, D., and Glazerman, S. (2004). "The effects of Teach for America on Students: Findings from a National Evaluation," Mathematica Policy Research Report, New York.

Figlio, D. and Getzler, L. (2006) "Accountability, Ability, and Disability: Gaming the System?" In T. Gronberg and D. Jansen (Eds), Advances in Applied Microeconomics, 14.

Figlio, D. and Kenny, L., (2007) "Individual teacher incentives and student performance," Journal of Public Economics, 91: 901-914.

Figlio, D. (2006) "Testing, Crime, and Punishment," Journal of Public Economics 90: 837-851.

Gibbons, R. (1998) "Incentives in Organizations," The Journal of Economic Perspectives 12(4): 115-132.

Glewwe, P., Ilias, N., and Kremer, M. (2003) "Teacher Incentives," NBER working paper #9671.

Hanushek, E. (2006) “School Resources” in E. Hanushek and F. Welch (Eds), Handbook of the Economics of Education.

Hanushek, E. and Rivkin, S. (2007) “Pay, Working Conditions, and Teacher Quality,” The Future of Children, 17(1): 69-86.

Herrmann, M., and Rockoff, J. (2009) "Work Disruption, Worker Health and Productivity: Evidence from Teaching," Columbia Business School.

Holmstrom, B. (1982) “Moral Hazard in Teams,” Bell Journal of Economics: 13, 324-340.

Holmstrom, B. and Milgrom, P. (1991) “Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design,” Journal of Law, Economics, and Organization, 7, Special Issue: Papers from the Conference on the New Science of Organization: 24-52.

Hoxby, C.M., and Leigh, A. (2005) “Pulled Away or Pushed Out? Explaining the Decline of Teacher Aptitude in the United States,” American Economic Review: 94, 236-240.

Itoh, H. (1991) “Incentives to Help in Multi-Agent Situations,” Econometrica: 59(3): 611-636.

Jackson, C. K. and Bruegmann, E. (2009) "Teaching Students and Teaching Each Other: The Importance of Peer Learning for Teachers," American Economic Journal: Applied Economics 1(4).

Jacob, B. (2005). “Accountability, Incentives and Behavior: Evidence from School Reform in Chicago,” Journal of Public Economics, 89: 761-796.

Jacob, B. and Levitt, S. (2003) "Rotten Apples: An Investigation Of The Prevalence And Predictors Of Teacher Cheating," The Quarterly Journal of Economics, 118(3): 843-877.

Kane, T., Rockoff, J., and Staiger, D. "What Does Certification Tell Us About Teacher Effectiveness? Evidence from New York City," Economics of Education Review, 27, no. 6 (2008): 615-631.

Lavy, V. (2002) “Evaluating the Effect of Teachers’ Group Performance Incentives on Pupil Achievement.” Journal of Political Economy, 110: 1286-1317.

Lavy, V. (2009) “Performance Pay and Teachers' Effort, Productivity and Grading Ethics,” American Economic Review, 99 (5): 1979-2011.

Lazear, E. (1986) “Salaries and Piece Rates,” Journal of Business, 59(3): 405-31.

Lazear, E. and Oyer, P. (2010) “Personnel Economics,” in Handbook of Organizational Economics, Gibbons, R. and Roberts, J. (eds).

Miller, R., Murnane R., and Willett J. (2008) “Do Worker Absences Affect Productivity? The Case of Teachers,” International Labour Review, 147 (1): 71–89.

Muralidharan, K., and Sundararaman, V. (2008) “Teacher Incentives in Developing Countries: Experimental Evidence from India,” 2008 Conference on Performance Incentives, Nashville, TN, National Center on Performance Incentives.

Oyer, P. and Schaefer, S. (2010) “Personnel Economics: Hiring and Incentives,” NBER working paper #15977.

Reback, R. (2008) "Teaching to the Rating: School Accountability and the Distribution of Student Achievement," Journal of Public Economics, 92: 1394–1415.

Rockoff, J., and Turner, L. J. (forthcoming) "Short Run Impacts of Accountability on School Quality," American Economic Journal: Economic Policy.

Winters, M. (2009) “The NYC Teacher Pay-for-Performance Program: Early Evidence from a Randomized Trial,” Manhattan Institute Civic Report No. 56.

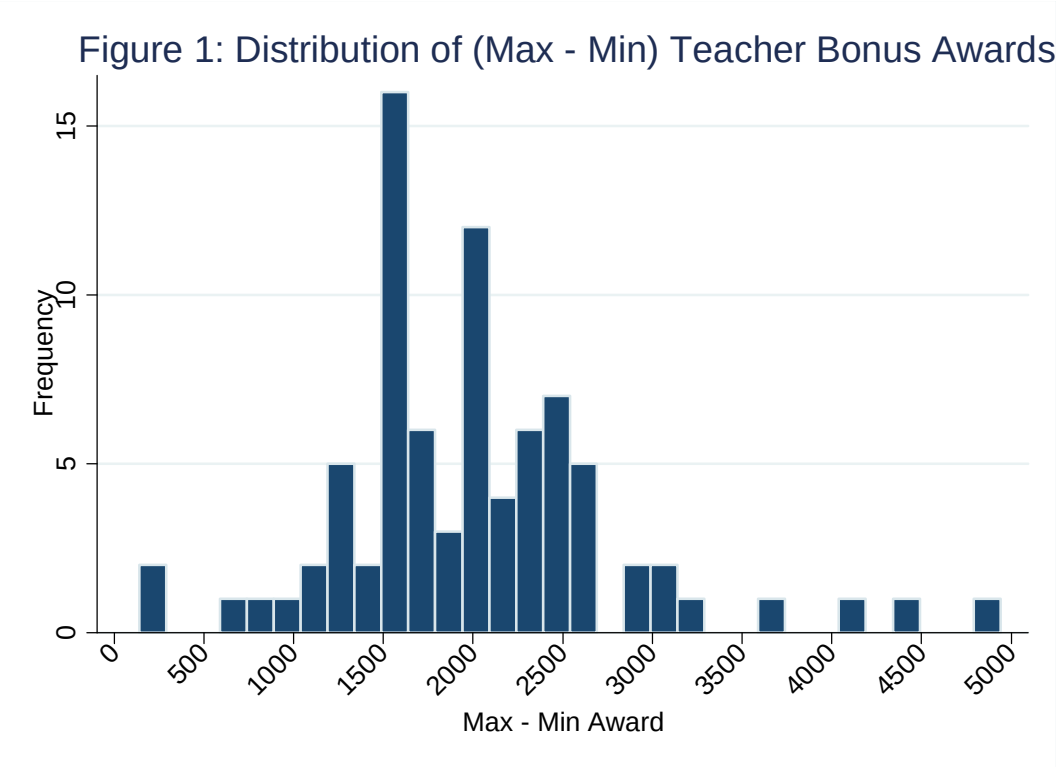
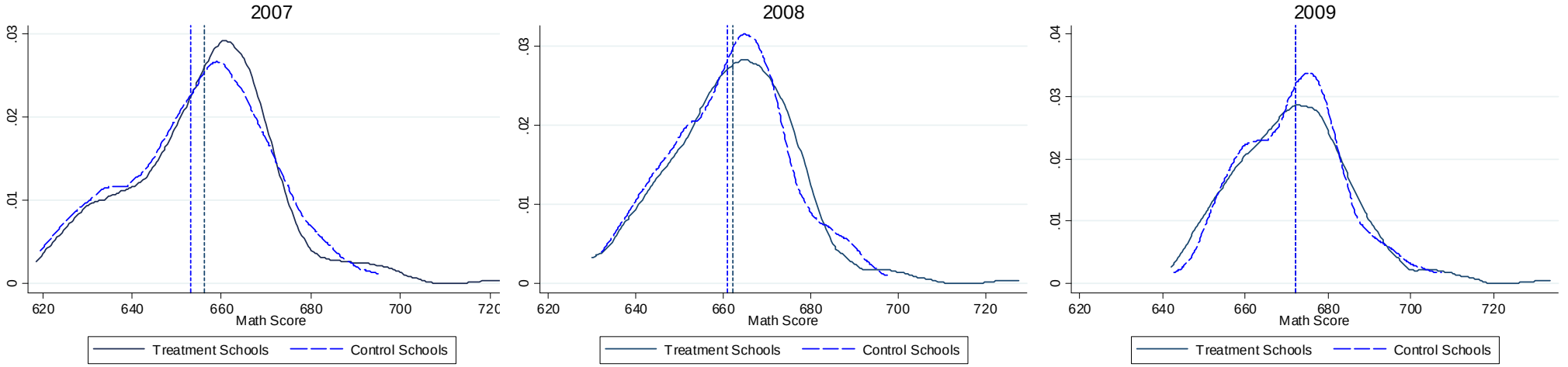


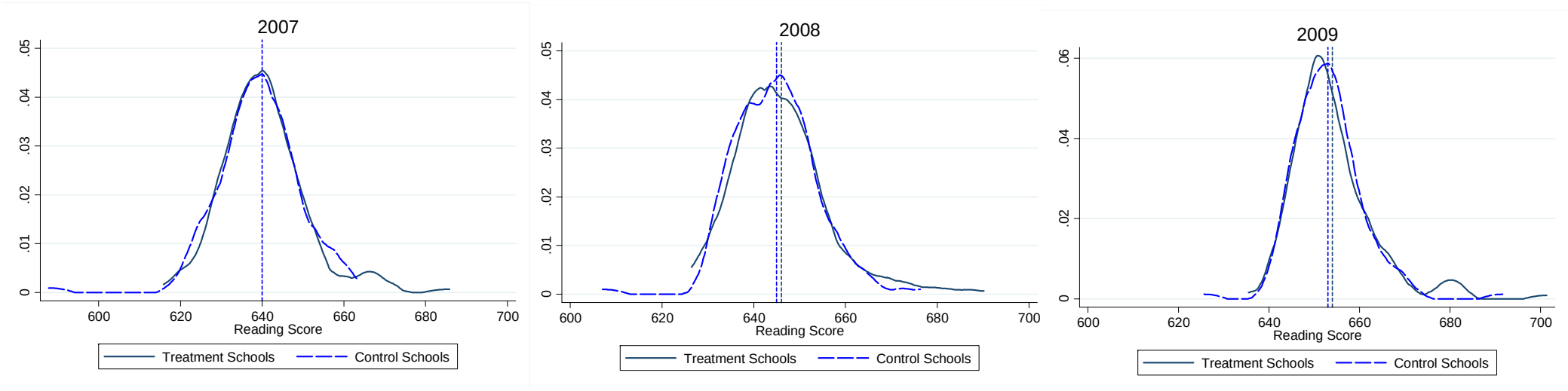
Figure 1: Distribution of (Max - Min) Teacher Bonus Awards
Frequency
0
5
10
15
0
500
1000
1500
2000
2500
3000
3500
4000
4500
5000
Max - Min Award

Figure 2: Distribution of Average Math Scores by Year and Treatment Status



Note: Dashed lines denote mean math scores for treatment and control schools.

Figure 3: Distribution of Average Reading Scores by Year and Treatment Status



Note: Dashed lines denote mean reading scores for treatment and control schools.

Table 1: Baseline School Characteristics by Original Assignment to Treatment and Control Groups

	Treatment Schools	Control Schools	Difference	p-value	Non-Experimental Schools
Number of Schools	181	128			614
Average enrollment	558	558	0	0.991	687
Average enrollment, tested grades	363	367	-4	0.852	459
Fraction elementary school	0.62	0.63	-0.01	0.912	0.63
Fraction middle school	0.26	0.27	-0.01	0.788	0.24
Fraction K-8 school	0.12	0.10	0.02	0.587	0.13
School Accountability Outcomes					
Peer index (mean = 0, sd = 1)	-0.91	-0.93	0.02	0.452	0.44
Overall accountability score	52.6	52.1	0.6	0.750	54.6
Target score	62.3	62.0	0.3	0.716	63.5
Student Characteristics					
Average math scale score (2007)	656	655	1	0.497	677
Change in math scale score (2006 to 2007)	10.6	10.3	0.3	0.717	8.8
Average reading scale score (2007)	640	640	1	0.603	660
Change in reading scale score (2006 to 2007)	1.5	2.0	-0.5	0.466	3.1
Fraction English Language Learner	0.19	0.19	0.01	0.614	0.11
Fraction special education	0.12	0.13	-0.01	0.246	0.09
Fraction free lunch	0.87	0.89	-0.02	0.315	0.62
Fraction Hispanic	0.56	0.53	0.03	0.428	0.33
Fraction Black	0.41	0.44	-0.03	0.425	0.30
Fraction White	0.01	0.01	0.00	0.640	0.20
Teacher Characteristics					
Number of teachers	55	55	0	0.952	60
Number of teachers, tested classrooms	13	13	0	0.572	14
Average years of experience	7.9	8.0	-0.1	0.703	8.6
Fraction with masters degree	0.48	0.47	0.02	0.579	0.45
Average absences/teacher (2007)	7.0	7.2	-0.2	0.447	3.7
Average absences/teacher, tested classrooms (2007)	7.2	7.4	-0.2	0.489	3.8
Fraction teachers not retained by DOE (2007)	0.11	0.12	0.00	0.513	0.09
Fraction teachers changing schools (2007)	0.07	0.07	0.01	0.321	0.04
Fraction of new teachers with MA	0.24	0.37	-0.13	0.407	0.45
Fraction of new teachers with prior experience	0.26	0.28	-0.02	0.493	0.39

Notes: Characteristics measured at beginning of 2007-2008 school year unless otherwise noted; average absences per teacher include absences taken for personal or self-treated sick leave.

Table 2: The Impact of Teacher Incentives on Average Absences/Teacher taken for Personal and Sick Leave, November 2007 - March 2008

	Sample Mean	(1) OLS	(2) OLS	(3) IV	(4) OLS	(5) OLS	(6) IV
A. All Teachers							
Treatment	3.82	0.100	0.018	0.022			
		(0.121)	(0.109)	(0.132)			
Observations		309	309	309			
B. Teachers with Tested Students							
Treatment	3.67	0.000	-0.078	-0.098	-0.155	-0.161	-0.181
		(0.177)	(0.171)	(0.210)	(0.178)	(0.177)	(0.215)
* Number of teachers with tested students *(mean = 0)*					0.048	0.030	0.042
					(0.018)*	(0.020)	(0.031)
Treatment effect point estimate:							
25th percentile (8 teachers)					-0.346	-0.283	-0.349
					(0.199)+	(0.197)	(0.237)
5th percentile (5 teachers)					-0.536	-0.404	-0.517
					(0.241)*	(0.242)+	(0.312)+
Observations		301	301	301	301	301	301
Additional covariates			X	X		X	X

Notes: + significant at 10%; * significant at 5%; ** significant at 1%; each column within Panels A and B denotes a separate regression; in columns (4) through (6) the number of teachers with tested students is demeaned; additional covariates include: prior (2006-2007) absences, school level, peer index, overall accountability score, percentage of students ELL, special education, free lunch recipients, and student race (African American and Hispanic); in column (3) and (6) regressions, actual treatment status is instrumented for with original treatment assignment; regressions are unweighted; schools with no teachers linked to tested students are dropped in Panel B regressions.

Table 3: Impact of Teacher Incentives on Student Math and Reading Achievement

	2007-2008				2008-2009			
	Mean (sd)	(1) OLS	(2) OLS	(3) IV	Mean (sd)	(4) OLS	(5) OLS	(6) IV
Reading	655	-0.876	-0.395	-0.486	662	-0.852	-0.384	-0.484
	(35)	(1.084)	(0.488)	(0.589)	(31)	(0.930)	(0.363)	(0.452)
Math	672	-1.418	-0.789	-0.970	680	-1.637	-0.705	-0.888
	(40)	(1.737)	(0.524)	(0.632)	(37)	(1.652)	(0.534)	(0.662)
Observations		309	309	309		305	302	302
Additional covariates			X	X			X	X

Notes: + significant at 10%; * significant at 5%; ** significant at 1%; scale score mean and sd for all NYC students, not schools, each cell denotes a separate regression; dependent variable: school average reading or math scale score; robust standard errors in parentheses; all regressions weighted by number of students tested in math or reading; additional covariates include: prior year scale score, indicators for school level, peer index, overall accountability score, percentage of students ELL, special education, free lunch recipients, and student race (African American and Hispanic); sample sizes differ across years due to the closure of four schools at the end of the 2007-2008 school year and the elimination of an additional three schools that did not receive 2008 accountability grades.

Table 4: Impact of Teacher Incentives on the Percentage of Students Achieving Proficiency

	2007-2008				2008-2009			
	Sample Mean	(1) OLS	(2) OLS	(3) IV	Sample Mean	(4) OLS	(5) OLS	(6) IV
Reading								
% Proficient	0.46	-0.020	-0.009	-0.011	0.58	-0.019	-0.006	-0.007
		(0.017)	(0.006)	(0.008)		(0.014)	(0.006)	(0.008)
Observations		309	309	309		305	302	302
Math								
% Proficient	0.67	-0.014	-0.009	-0.012	0.77	-0.018	-0.010	-0.012
		(0.020)	(0.007)	(0.008)		(0.016)	(0.005)+	(0.007)+
Observations		309	309	309		305	302	302
Additional covariates			X	X			X	X

Notes: + significant at 10%; * significant at 5%; ** significant at 1%; each cell denotes a separate regression; dependent variable: % proficient in math or in reading; robust standard errors in parentheses; all regressions weighted by number of students tested in math or reading; additional covariates include: prior year scale score, indicators for school level, peer index, overall accountability score, percentage of students ELL, special education, free lunch recipients, and student race (African American and Hispanic); sample sizes differ across years due to the closure of four schools at the end of the 2007-2008 school year and the elimination of three schools that did not receive 2008 accountability grades.

Table 5: Free-riding and the Impact of Teacher Incentives on Student Math and Reading Achievement

	Reading			Math		
	(1) OLS	(2) OLS	(3) IV	(4) OLS	(5) OLS	(6) IV
A. Main Results						
Treatment	-0.876	-0.395	-0.486	-1.418	-0.789	-0.970
	(1.084)	(0.488)	(0.589)	(1.737)	(0.524)	(0.632)
Observations	309	309	309	309	309	309
B. Interactions - Number of Teachers						
Treatment	-0.327	-0.160	-0.241	-0.007	-0.510	-0.652
	(1.124)	(0.505)	(0.595)	(1.766)	(0.565)	(0.657)
* Number of reading or math teachers	-0.339	-0.212	-0.311	-0.822	-0.213	-0.358
(mean = 0)	(0.207)	(0.114)+	(0.164)+	(0.328)*	(0.119)+	(0.214)+
Treatment effect point estimate:						
25th percentile (8 teachers)	1.028	0.688	1.004	3.282	0.344	0.780
	(1.569)	(0.768)	(0.768)	(2.437)	(0.861)	(0.861)
5th percentile (5 teachers)	2.383	1.537	2.249	6.571	1.198	2.212
	(2.243)	(1.157)	(1.532)	(3.492)+	(1.271)	(1.950)
Observations	303	303	303	303	303	303
Additional controls		X	X		X	X

Notes: + significant at 10%; * significant at 5%; ** significant at 1%; each column denotes a separate regression; Panel B measures of the number of reading/math teachers are demeaned; additional controls include: prior (2007) school test score, school level, peer index, overall accountability score, percentage of students ELL, special education, free lunch recipients, and student race (African American and Hispanic); regressions are weighted by number of tested students; in column (3) and (6) regressions, actual treatment status is instrumented for with original treatment assignment; schools with no teachers linked to tested students are dropped.

Table 6: Impact of Teacher Incentives on Student and Teacher Survey Outcomes

	(1) OLS	(3) OLS	(4) IV
A. Student Survey Outcomes			
Essays and Projects	0.120	0.231	0.295
	(0.163)	(0.153)	(0.190)
Group & Hands-on Learning Activities	0.182	0.163	0.207
	(0.194)	(0.193)	(0.238)
Tutoring Offered Before/After School	-0.088	0.191	0.241
	(0.188)	(0.166)	(0.200)
Observations	112	112	112
B. Teacher Survey Outcomes			
Use of Student Data	-0.075	-0.072	-0.089
	(0.108)	(0.106)	(0.128)
High Expectations For Students	-0.134	-0.120	-0.147
	(0.103)	(0.094)	(0.113)
Observations	305	305	305
Additional controls		X	X

Notes: + significant at 10%; * significant at 5%; ** significant at 1%; robust standard errors in parentheses; each cell denotes a separate regression; all regressions control for prior (2007) survey outcome; additional controls include: school level, peer index, overall accountability score, percentage of students ELL, special education, free lunch recipients, and student race (African American and Hispanic), all regressions weighted by number of survey respondents; in column (3) regressions, actual treatment status is instrumented for with original treatment assignment; see text for description of survey measures.

Table 7: The Impact of Teacher Incentives on Teacher Turnover and the Qualifications of New Teachers

	Sample Mean	(1) OLS	(2) OLS	(3) IV
A. Teacher Turnover, 2008-2009				
Fraction of teachers not retained by school district	0.11	0.003	0.005	0.006
		(0.006)	(0.006)	(0.007)
Fraction of teachers leaving for another NYC school	0.07	0.007	0.006	0.007
		(0.006)	(0.005)	(0.007)
Observations		302	302	302
B. Characteristics of New Teachers, 2009				
Fraction of new teachers from Teach for America program	0.11	-0.005	-0.008	-0.009
		(0.038)	(0.030)	(0.034)
Fraction of new teachers with MA	0.25	0.006	0.014	0.016
		(0.032)	(0.032)	(0.036)
Fraction of new teachers with prior teaching experience	0.21	0.014	0.029	0.034
		(0.028)	(0.027)	(0.031)
Observations		267	257	257
Additional covariates			X	X

Notes: + significant at 10%; * significant at 5%; ** significant at 1%; robust standard errors in parentheses; each cell denotes a separate regression; additional covariates include: prior (2007-2008) fraction of teachers not retained or fraction of teachers leaving for another school (Panel A), prior fraction of teachers with MA or prior experience (Panel B); school level, peer index, overall accountability score, percentage of students ELL, special education, free lunch recipients, and student race (African American and Hispanic); all regressions weighted by number of teachers (panel A) or number of new teachers (panel B); schools without new teacher hires dropped from Panel B regressions; in column (3) regressions, actual treatment status is instrumented for with original treatment assignment.

Table A1: Baseline School Characteristics by Participation Vote

	Voted "yes"	Voted "no"	Difference	p-value
Number of Schools	154	25		
Average enrollment	556	574	-17	0.736
Average enrollment, tested grades	363	361	2	0.978
Fraction elementary school	61%	72%	-11%	0.296
Fraction middle school	12%	8%	4%	0.535
Fraction K-8 school	27%	20%	7%	0.485
School Accountability Outcomes				
Peer index (mean = 0, sd = 1)	-0.91	-0.87	-0.05	0.250
Overall accountability score	52.3	55.1	-2.8	0.419
Target score	62.1	64.2	-2.1	0.309
Student Characteristics				
Average math scale score (2007)	655	661	-6	0.111
Change in math scale score (2006 to 2007)	10.7	10.2	0.4	0.742
Average reading scale score (2007)	640	644	-4	0.047
Change in reading scale score (2006 to 2007)	1.6	0.2	1.4	0.359
Fraction English Language Learner	20%	18%	2%	0.540
Fraction special education	12%	12%	0%	0.756
Fraction free lunch	88%	86%	2%	0.645
Fraction Hispanic	56%	54%	3%	0.640
Fraction Black	41%	42%	-1%	0.833
Fraction White	1%	1%	0%	0.777
Teacher Characteristics				
Number of teachers	54	56	-2	0.685
Number of teachers, tested classrooms	12	13	-1	0.408
Average years of experience	7.6	7.8	-0.1	0.161
Fraction with masters degree	34%	40%	-6%	0.786
Average absences (2007)	4.1	4.0	0.1	0.433
Average absences, tested classrooms (2007)	4.2	4.3	-0.1	0.977
Fraction teachers not retained by DOE (2007)	12%	10%	2%	0.224
Fraction teachers changing schools (2007)	7%	7%	0%	0.980
Fraction of new teachers with MA	34%	40%	-6%	0.285
Fraction of new teachers with prior experience	24%	38%	-14%	0.005

Notes: Characteristics measured at beginning of 2007-2008 school year unless otherwise noted; + difference between treatment and control significant at 10%, * 5%, ** 1%; average absences measured between November 2006 and March 2007.

Table A2: The Effect of Teacher Incentives on the Percentage and Composition of Tested Students, 2007 - 2008

	(1) OLS	(2) OLS	(3) IV
Math			
Percentage of Students Tested	0.001	0.000	0.000
	(0.004)	(0.003)	(0.004)
Observations	309	309	309
Percentage of tested students ELL	0.008	-0.000	-0.000
	(0.016)	(0.001)	(0.002)
Observations	260	260	260
Percentage of tested students special education	-0.002	0.000	0.000
	(0.006)	(0.003)	(0.003)
Observations	294	294	294
Reading			
Percentage of Students Tested	-0.001	-0.003	-0.003
	(0.005)	(0.003)	(0.004)
Observations	309	309	309
Percentage of tested students ELL	0.010	0.002	0.003
	(0.016)	(0.002)	(0.002)
Observations	257	257	257
Percentage of tested students special education	-0.000	0.002	0.002
	(0.006)	(0.003)	(0.003)
Observations	295	295	295
Additional covariates		X	X

Notes: + significant at 10%; * significant at 5%; ** significant at 1%; robust standard errors in parentheses; each cell denotes a separate regression; all regressions control for prior (2007) percentage of students tested, percentage ELL, or percentage special education; additional controls include: school level, peer index, overall accountability score, percentage of students ELL, special education, free lunch recipients, and student race (African American and Hispanic), regressions of percentage of students tested weighted by total enrollment, all other regressions weighted by number of tested students; in column (3) regressions, actual treatment status is instrumented for with original

Table A3: Heterogeneity in Impact of Teacher Incentives on Student Math and Reading Achievement by Accountability Grade

	2007-2008			2008-2009		
	(1) OLS	(2) OLS	(3) IV	(4) OLS	(5) OLS	(6) IV
A. Reading						
Treatment*D or F	-3.719	0.175	0.188	-3.533	-1.924	-3.246
	(2.147)+	(1.116)	(1.244)	(2.321)	(1.300)	(2.178)
Treatment*B or C	0.454	-0.733	-0.892	-0.488	-0.379	-0.465
	(1.309)	(0.615)	(0.726)	(0.976)	(0.435)	(0.518)
Treatment* A	-1.856	0.063	0.091	1.179	-0.091	-0.113
	(2.489)	(1.122)	(1.439)	(1.802)	(0.686)	(0.843)
Test A/B = C = D/F (pvalue)	0.231	0.698	0.685	0.277	0.450	0.400
Observations	309	309	309	305	302	302
B. Math						
Treatment*D or F	-5.292	-0.329	-0.379	-7.273	-3.326	-5.582
	(3.407)	(1.144)	(1.279)	(4.017)+	(2.388)	(3.924)
Treatment*B or C	0.944	-0.400	-0.470	-0.686	-0.463	-0.574
	(2.163)	(0.700)	(0.825)	(1.888)	(0.658)	(0.786)
Treatment* A	-3.703	-1.542	-2.031	1.511	-0.603	-0.749
	(3.608)	(1.084)	(1.447)	(2.726)	(0.952)	(1.174)
Test A/B = C = D/F (pvalue)	0.238	0.653	0.625	0.193	0.515	0.460
Observations	309	309	309	305	302	302
Additional covariates		X	X		X	X

Notes: + significant at 10%; * significant at 5%; ** significant at 1%; each column within a panel denotes a separate regression; dependent variable: school average reading or math scale score interacted with indicator for school grade; robust standard errors in parentheses; all regressions weighted by number of students tested in math or reading; additional covariates include: prior year scale score, indicators for school level, peer index, overall accountability score, percentage of students ELL, special education, free lunch recipients, and student race (African American and Hispanic); sample sizes differ across years due to the closure of four schools at the end of the 2007-2008 school year and the elimination of an additional three schools that did not receive 2008 accountability grades.

Table A4: Heterogeneity in Impact of Teacher Incentives on Student Math and Reading Achievement by Bonus Receipt

	2008-2009		
	(1) OLS	(2) OLS	(3) IV
A. Reading			
Treatment	-0.759	-0.431	-0.568
	(0.992)	(0.447)	(0.680)
* Any Bonus (predicted)	0.944	0.287	0.386
	(1.873)	(0.741)	(1.133)
Observations	302	302	302
B. Math			
Treatment	-0.804	-0.673	-0.684
	(1.899)	(0.690)	(1.052)
* Any Bonus (predicted)	0.269	0.047	-0.148
	(3.238)	(1.080)	(1.662)
Observations	302	302	302
Additional covariates		X	X

Notes: + significant at 10%; * significant at 5%; ** significant at 1%; each column within a panel denotes a separate regression; dependent variable: school average reading or math scale score interacted with indicator for school grade; robust standard errors in parentheses; all regressions weighted by number of students tested in math or reading; additional covariates include: prior year scale score, indicators for school level, peer index, overall accountability score, percentage of students ELL, special education, free lunch recipients, and student race (African American and Hispanic); in column (3), actual treatment status is instrumented for with original treatment assignment